The present invention achieves this objective with a novel semantic based method of identifying records of interest based on the similarity of their content to the meaning of the input phrase. In accordance with the invention, "expert knowledge" of the content of the database is stored in a computer file. This file's architecture allows a computer program to supplement a user's input with additional information that expresses the meaning of the request more fully in the context of the database. The invention also employs a novel search technique that rates the similarity of each database record to the meaning of the user request. While the resulting search engine accommodates unformatted, a natural language input, it is not dependent on the use of precise terminology. Further, since its fundamental record identification function is based on semantic similarity rather than exact character string matching, the search techniques can tolerate partially incorrect user input.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a block diagram illustrating the modules of the present invention and how they relate to each other in operation.

Figure 2 is a flow chart that illustrates the steps performed to identify the core vocabulary of a database.

Figure 3 is a flow chart that illustrates the steps performed to construct a predominate semantic structure that effectively models the database content.

Figure 4 is a flow chart that illustrates the steps performed to associate the core vocabulary within the predominate semantic structure.

Figure 5 is a flow chart that illustrates the steps performed to supplement the core vocabulary and capture the contextual significance of the usage of each term.

Figure 6 is a flow chart that illustrates the steps performed to interpret the meaning of a user request.

Figure 7 is a flow chart that illustrates the steps performed to determine the similarity of a database record to the meaning of a user request.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides a search methodology that identifies records in a specialized database that have content that is similar to the meaning of a user request.

Figure 1 provides an overview of the invention's process. A sophisticated user of the subject database (the "domain expert") is presented with computer generated characteristics of the database, along with a number of possible organizational templates. The domain expert then constructs an appropriate semantic organizational structure for the content of the database. The expert also supplements the database's core vocabulary and assigns all terms within the semantic structure, thereby incorporating his domain expertise

into the Lexicon file. The information in the Lexicon file is used to supplement a user request, to more fully express it's meaning within the context of the database. The expanded query is then used to rate the similarity of the content of each database record to the meaning of the user request. Entries with high similarity are presented to the user for subjective review.

Figure 2 illustrates how the invention implements Praeto's Principle (the so called "80/20 rule) to identify the database's core vocabulary. The computer program performs a word usage distribution analysis on the entire text of the database, identifying the total number of times each word is used. The computer program then sorts the words in descending order of usage and prepares a matrix that associates the number of times a word is used with the cumulative number of words in the rank ordering prior to that word. The computer program then identifies the first point of inflection of the associated curve by using the technique of Newton's Approximation to identify the first significant local minimum of the second derivative of usage with respect to the cumulative number of words. The computer program then identifies the core vocabulary of the database as the set of words in the matrix prior to the point of inflection.

Figure 3 illustrates how the invention captures the predominate semantic structure of the database. The computer generates a random sample of descriptions from the database that is statistically representative of the population at a 95% confidence level. These descriptions are presented to a domain expert along with a set of possible semantic organizational templates (i.e. potential conceptual groupings of information such as color,